The Report Committee for Daniel Lewis Mitchell
Certifies that this is the approved version of the following report:

Spatial Interpolation with Gaussian Processes and Spatially Varying Regression Coefficients

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Timothy H. Keitt

James G. Scott

Spatial Interpolation with Gaussian Processes and Spatially Varying Regression Coefficients

by

Daniel Lewis Mitchell, B.S.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

August 2015

Dedication

To my wife, Jamie, and our children, Laudon and Weston.

Acknowledgements

Foremost I would like to thank my advisor, Timothy Keitt, for believing in me during good times and bad. I would also like to thank my co-advisor, James Scott, for teaching me so much. And I would like to thank Vicki Keller for all her encouragement and assistance. Finally, I would like to thank my family for their support.

Abstract

Spatial Interpolation with Gaussian Processes and Spatially Varying Regression Coefficients

Daniel Lewis Mitchell, M.S. Stat

The University of Texas at Austin, 2015

Supervisor: Timothy H. Keitt

Linear regression is undoubtedly one of the most widely used statistical techniques, however because it assumes independent observations it can miss important features of a dataset when observations are spatially dependent. This report presents the spatially varying coefficients model, which augments a linear regression with a multivariate Gaussian spatial process to allow regression coefficients to vary over the spatial domain of interest. We develop the mathematics of Gaussian processes and illustrate their use, and demonstrate the spatially varying coefficients model on simulated data. We show that it achieves lower prediction error and a better fit to data than a standard linear regression.

Table of Contents

List of Tables

List of Figures

Introduction

Linear regression is undoubtedly one of the most widely used statistical techniques, however because it assumes independent observations it can miss important features of a dataset when observations are in some way dependent. In this report we discuss a technique for extending linear regression to point-referenced data, which we anticipate will exhibit spatial dependence. Generally we expect observations recorded at nearby locations to be more alike than observations recorded at distant locations. Such dependency isn't captured by standard linear regression. This report presents a technique known as *spatially varying coefficients* that augments a linear regression with a spatial process to model a spatially dependent response.

The spatially varying coefficients model replaces the fixed coefficients of a linear regression with Gaussian processes conditioned on the observed data, effectively resulting in a regression in which the coefficients vary over the spatial domain. Because the regression coefficients are permitted to vary over the spatial domain of interest, the model is able to capture changing relationships between the response variable and its covariates. This enables us to make more accurate predictions and create maps showing how the coefficients change in space.

In this report we will develop the fundamentals of Gaussian processes and illustrate how they can be used for prediction. We will also show how their properties can be inferred from data. Then we will introduce the spatially varying coefficients model and demonstrate its use with a case study on simulated data.

Gaussian Processes

A Gaussian process is a collection of random variables $\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$, any finite subset of which has a joint multivariate normal distribution. The set $\mathcal{X}$ is an arbitrary index set such as the integers in a discrete context or the real numbers in a continuous context. An observation of a Gaussian process is called a realization of the process. Although the index set may be infinite as in the case of the real numbers, in practice only a finite number of the random variables are ever observed, yielding a partial realization of the process. If the index set is finite, then we have a Gaussian distribution rather than a process.

The usual objective of a Gaussian process analysis is to determine the joint distribution of a set of unobserved variables in $\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ conditioned on the known values of some number of observed variables. With the joint distribution available it is possible to make predictions regarding the unobserved variables. To be concrete, suppose the index set $\mathcal{X}$ consists of spatial locations, for example latitude and longitude coordinates. Given a partial realization of a process at a set of locations $\{\boldsymbol{x}_i : i = 1, \ldots, N\} \subset \mathcal{X}$, for example measurements of atmospheric conditions such as temperature or pollutants recorded at N monitoring sites, one seeks to make predictions and obtain the distribution of atmospheric conditions at some unmonitored location $\boldsymbol{x}^* \in \mathcal{X}$.

Because a normal random variable is fully characterized by its first two moments, a Gaussian processes is fully characterized by a mean function $m : \mathcal{X} \to \mathbb{R}$ and a

covariance function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ in the sense that $E(\boldsymbol{x}) = m(\boldsymbol{x})$ and $\text{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2) = C(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Covariance functions will be examined in later sections. Although it isn't necessary, typically the mean function m is taken be identically zero and the process mean is modeled separately, for example by a regression or a generalized additive model. Then the Gaussian process is applied to the differences between the observations and the mean.

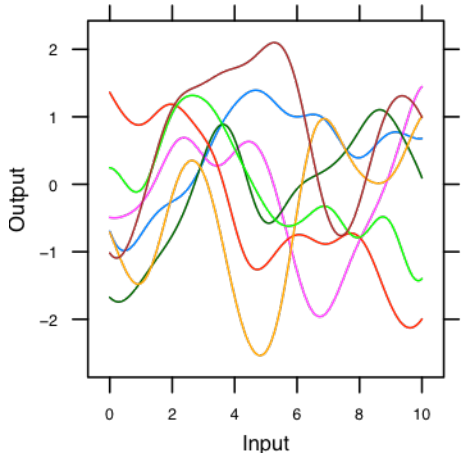

Figure 1: Sample realizations of a mean zero Gaussian process with squared exponential covariance function. These can be regarded as random functions with behavior determined by the covariance.

Gaussian processes are widely used in statistical applications such as nonlinear regression and machine learning applications such as classification. In a Bayesian context they often serve as priors on an unknown function $f : \mathbb{R} \to \mathbb{R}$, with the index set $\mathcal{X}$ being the real numbers. From this point of view, a Gaussian process is regarded as a random function where for each $x \in \mathbb{R}$ there exists a normally distributed random

variable $f(x)$ with a mean and variance determined by the properties of the process. This is indicated by the notation $f \sim GP(m, C)$. Figure 1 shows several random functions drawn from a Gaussian process.

The indices $\mathcal{X}$ are referred to as inputs while the random variables $\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ are referred to as outputs. An observed set of inputs and outputs is referred to as training data while a set of inputs with unobserved outputs is referred to as test data. Throughout this section, test data is denoted with an asterisk.

THE MULTIVARIATE NORMAL DISTRIBUTION

Because of its central role in Gaussian process analysis, it is useful to review some properties of the multivariate normal distribution. In particular, the marginal and especially the conditional distributions will be heavily used. Suppose that the random vector $\boldsymbol{y}$ has a $p > 1$ dimensional multivariate normal distribution, so that $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \Sigma)$. The probability density is

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right).$$

We can partition $\boldsymbol{y}$ into two sets of components, $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)^T$, where $\boldsymbol{y}_1$ has dimension q, $0 < q < p$, and $\boldsymbol{y_2}$ has dimension $p - q$. The mean vector $\boldsymbol{\mu}$ is likewise partitioned as $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^T$, as is the covariance matrix, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, with the block Σ_{12} having dimensions $q \times (p - q)$. This partitioning is expressed symbolically as

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

Marginal and Conditional Distributions

With $\boldsymbol{y}$ partitioned as above, the marginal distribution of $\boldsymbol{y}_1$ is obtained simply by deleting all entries related to the second component $\boldsymbol{y}_2$. Thus the marginal distribution is $\boldsymbol{y}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$. Similarly the marginal distribution of $\boldsymbol{y}_2$ is $N(\boldsymbol{\mu}_2, \Sigma_{22})$. These facts will be used in the discussion of covariance functions and dimension reduction in later sections.

If component $\boldsymbol{y}_2$ has been observed, then the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is

$$\boldsymbol{y}_1 \mid \boldsymbol{y}_2 \sim N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

The conditional distribution in particular is the workhorse of Gaussian process analysis and Bayesian inference in general. As we shall see, the conditional distribution will allow us to obtain the distribution of unobserved variables of interest conditioned on the variables that have been observed. In other words, it allows us to obtain the distribution of what we would like to know given what we actually know.

GAUSSIAN PROCESS PREDICTIVE DISTRIBUTIONS

Suppose that a partial realization $\boldsymbol{y}$ of a mean zero Gaussian process with covariance function C has been observed at a set X of n locations and we wish to make predictions at a set X^* of n^* unobserved locations. To make predictions we need to obtain the distribution of the test outputs $\boldsymbol{y}^*$ given the training outputs $\boldsymbol{y}$. This simply means finding the conditional distribution of the unobserved outputs $\boldsymbol{y}^*$ conditioned on the observed outputs $\boldsymbol{y}$.

The Case of Noise-free Observations

If a mean zero process is observed exactly, then the observations are modeled as

$$y_i = f(\boldsymbol{x}_i)$$

$$f \sim GP(0, C).$$

In this case the training data consists of pairs $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$. The joint distribution of the training outputs and test outputs is an $n + n^*$ dimensional multivariate normal distribution, which we partition into two components:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y}^* \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix} C(X,X) & C(X,X^*) \\ C(X^*,X) & C(X^*,X^*) \end{pmatrix}\right).$$

Here $C(X, X^*)$ is the $n \times n^*$ matrix where element (i, j), $1 \leq i \leq n$ and $1 \leq j \leq n^*$, is given by $C(\boldsymbol{x}_i, \boldsymbol{x}_j^*)$, the covariance of the i-th training output and the j-th test output. The matrices $C(X, X)$ and $C(X^*, X^*)$ are defined similarly, and of course $C(X^*, X) = C(X, X^*)^T$ by symmetry. Let us introduce a convenient shorthand form of notation: let $C_{nn} = C(X, X)$, $C_{n*} = C(X, X^*)$, $C_{*n} = C(X^*, X)$, and $C_{**} = C(X^*, X^*)$. To determine the conditional distribution of $\boldsymbol{y}^*$ given $\boldsymbol{y}$ we apply the earlier result on multivariate normal conditionals, obtaining

$$\boldsymbol{y}^* \mid \boldsymbol{y} \sim N(C_{*n} C_{nn}^{-1} \boldsymbol{y}, C_{**} - C_{*n} C_{nn}^{-1} C_{n*}).$$

It's worth noting that the mean is a linear combination of the observed outputs $\boldsymbol{y}$ but the covariance does not depend on $\boldsymbol{y}$. This latter property is perhaps undesirable from a modeling perspective and may justify the use of other kinds of processes, such as a Student-t process.

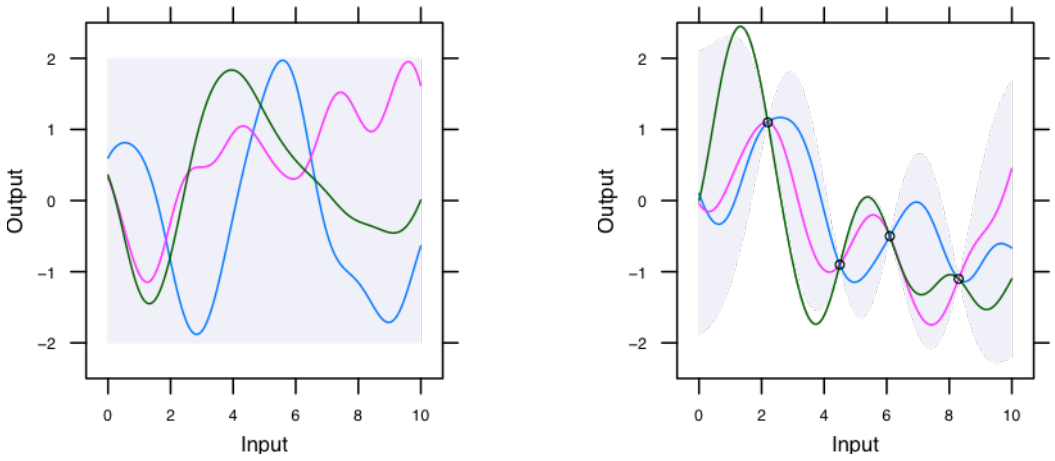

Figure 2: The left panel shows three samples from a mean zero Gaussian process prior with squared exponential covariance function, as well as a pointwise confidence band of plus and two standard deviations. The right panel shows three samples from the posterior process after observing four training points without error. The samples pass exactly through the training data as a consequence of observing the process without error. The confidence bands shrink to zero width at the training points for the same reason.

Figure 2 shows an example of process prediction when the process is observed exactly. The left panel shows three sample functions drawn from a Gaussian process prior, along with a pointwise confidence band of plus and minus two standard deviations. The right panel shows three sample functions drawn from the posterior after observing four training points. Notice that each of these functions is constrained to pass exactly through the training data, a consequence of the process being observed without error. This constraint is reflected in the confidence band, which shrinks to zero width at the training points.

The Case of Noisy Observations

If the process has been observed with independent Gaussian error rather than exactly, then the observations are modeled as

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

$$f \sim GP(0, C)$$

$$\epsilon \sim N(0, \sigma^2).$$

Again the training data consists of pairs $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$. Typically it is the process f that is of scientific interest, not the noisy version of the process y, so we seek the joint distribution of the training outputs and f at the test points:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{f}^* \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} C_{nn} + \sigma^2 I_n & C_{n*} \\ C_{*n} & C_{**} \end{pmatrix}\right).$$

In this case the conditional distribution of $\boldsymbol{f}^*$ given $\boldsymbol{y}$ is

$$\boldsymbol{f}^* \mid \boldsymbol{y} \sim N(C_{*n}(C_{nn} + \sigma^2 I_n)^{-1}\boldsymbol{y}, C_{**} - C_{*n}(C_{nn} + \sigma^2 I_n)^{-1}C_{n*}).$$

The only difference between this conditional and the noise-free conditional is the addition of the observation error variance matrix $\sigma^2 I_n$ in the inverse terms, which is diagonal by the assumption of independent error. If we wish to predict the noisy process y^* at the test points, we simply add the observation error variance $\sigma^2 I_{n^*}$ to the above covariance matrix, obtaining

$$\boldsymbol{y}^* \mid \boldsymbol{y} \sim N(C_{*n}(C_{nn} + \sigma^2 I_n)^{-1}\boldsymbol{y}, C_{**} - C_{*n}(C_{nn} + \sigma^2 I_n)^{-1}C_{n*} + \sigma^2 I_{n^*}).$$

The error variance σ^2 is often referred to as the nugget in spatial statistics literature. It captures measurement error and micro-scale spatial variation. Sometimes a nugget term is

included in the covariance function, but it is unidentifiable under noisy sampling. This is discussed in more detail in the next section.

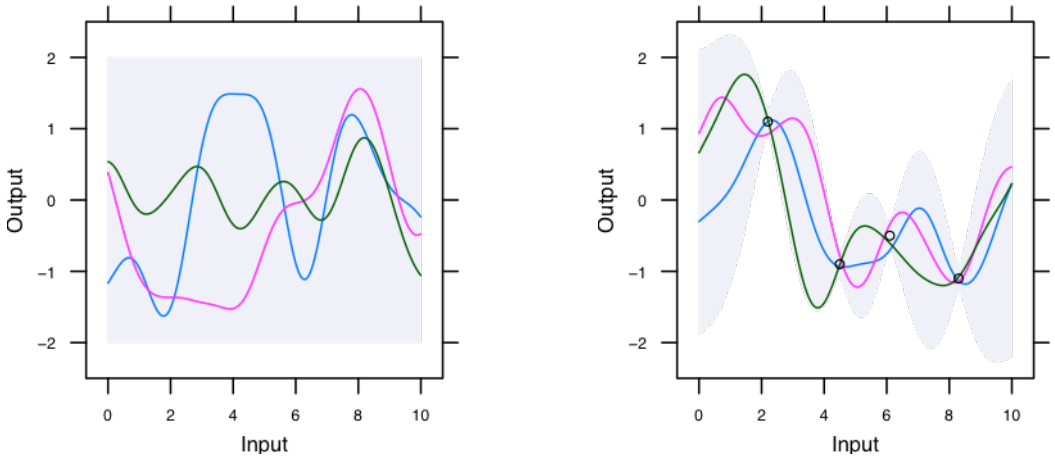

Figure 3: The left panel is the same as the left panel of Figure 2. The right panel shows three samples from the posterior process after observing four training points with Gaussian error. In contrast to Figure 2, the samples are not constrained to pass exactly through the training data and the confidence bands no longer shrink to zero width at the training points.

Figure 3 shows the prediction results. The left panel is the same as Figure 2, since the prior is the same in both cases. The right panel shows that under noisy observation the posterior draws are no longer constrained to pass exactly through the training data. Instead the inclusion of observation error introduces some freedom around the observed points, which is reflected in the nonzero width of the confidence bands at the training points.

COVARIANCE FUNCTIONS

In both the noise-free case and the noisy observation case the covariance function plays a key role in process prediction. The predicted mean is a linear combination of the observed values with coefficients determined by the covariance function, and the predicted covariance is determined by the covariance function. The smoothness of the process is also determined by the properties of the covariance function. In general, the covariance function captures the idea that observations at nearby points will tend to be more alike than observations at far away points. What is meant by nearby depends on the application, but in spatial statistics it refers to Euclidean distance or great circle distance. In classification problems it usually refers to how similar cases are, or stated differently it refers to distance in feature space.

In general, an arbitrary function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ will not be a valid covariance function. To be a valid covariance function, it must give rise to symmetric positive semi-definite covariance matrices for all finite sets of input points in $\mathcal{X}$. For all sets $\{\boldsymbol{x}_i : i = 1, \ldots, n\} \subset \mathcal{X}$, for all $n > 0$, the matrix Σ with $\Sigma_{ij} = C(\boldsymbol{x}_i, \boldsymbol{x}_j)$ must be symmetric positive semi-definite.

If $C(\boldsymbol{x}_1, \boldsymbol{x}_2)$ depends on $\boldsymbol{x_1}$ and $\boldsymbol{x}_2$ only through $\boldsymbol{x}_1 - \boldsymbol{x}_2$ then the spatial process is said to be stationary; in this case the covariance is invariant to translation, which is to say in a spatial context that the covariance is independent of location. If furthermore the covariance depends only on the Euclidean distance $|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ then the process is said to be isotropic; in this case the covariance is invariant to both translation and rotation, which is to say it is independent of both location and direction. Whether or not these are

realistic properties depends on the modeling context. In any case they are convenient properties and serve as the foundation of more complex nonstationary, nonisotropic models.

Usually parametric covariance functions are used which allow various properties of the process to be controlled or learned from data. In the next section we will discuss a popular covariance function and compare some of its important special cases.

The Matérn Class

A widely used covariance function is the Matérn covariance function, defined as

$$C(\boldsymbol{x}_1, \boldsymbol{x}_2; \nu, \sigma^2, r) = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}\, |\boldsymbol{x}_1 - \boldsymbol{x}_2|}{r} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}\, |\boldsymbol{x}_1 - \boldsymbol{x}_2|}{r} \right),$$

with $\nu > 0$ and $r > 0$. Here $\Gamma(\nu)$ is the gamma function and K_{ν} is a modified Bessel function of the second kind. Despite its somewhat complicated definition in terms of special functions, this is a popular choice of covariance function because the parameter ν explicitly controls the smoothness of the resulting process, with larger values of ν producing smoother processes. For this reason the Matérn covariance function is said to define a class or family of covariance functions. When ν is a half-integer, that is when $\nu = n + 1/2$ for a nonnegative integer n, then the Matérn covariance factors into a simple form involving an exponential term and a polynomial of degree n.

One important special case of the Matérn covariance is when $\nu = 1/2$, giving the exponential covariance function

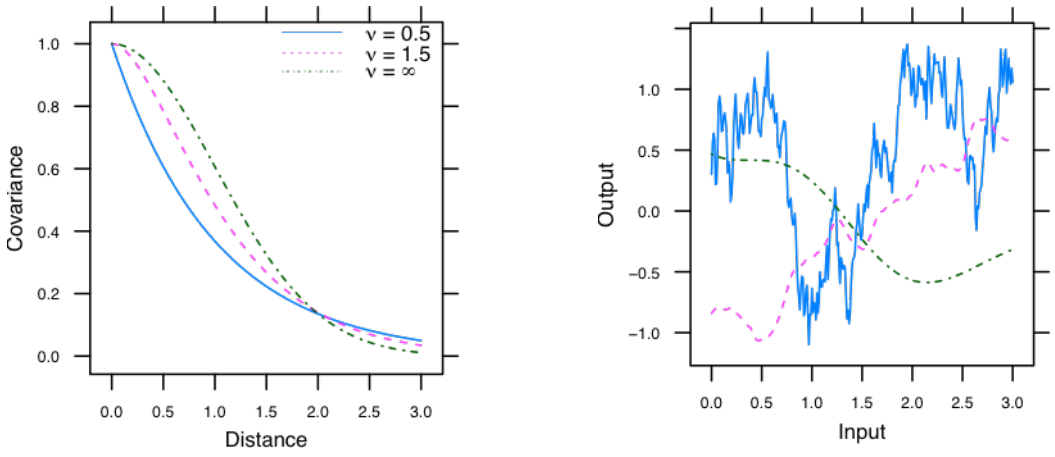

Figure 4: The left panel shows three Matérn class covariance functions with different values of the smoothness parameter ν. The right panel shows a sample realization from each.

$$C(\boldsymbol{x}_1, \boldsymbol{x}_2; \sigma^2, r) = \sigma^2 \exp\left(-\frac{|\boldsymbol{x}_1 - \boldsymbol{x}_2|}{r}\right).$$

In the spatial statistics literature $\phi = 1/r$ is called the spatial decay parameter and determines the range of the spatial process. The range is defined as the distance $|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ at which the covariance falls to zero. When the covariance reaches zero only in the limit as distance approaches infinity, as it does in the case of the exponential covariance, there is a related concept known as the practical range. The practical range is the distance at which the covariance becomes very small compared to its value at zero distance, usually when it has decreased by 95%. For the exponential covariance the practical range is $-\log(0.05)\, r \approx 3r$. We will illustrate the effect of the range parameter later in a section on learning covariance parameters. The marginal variance parameter σ^2 controls the overall variability of the process. It is known as the partial sill in geostatistics, and is

defined as the difference between the sill and the nugget. The sill is the limit of the covariance function as distance approaches zero.

Another important special case of the Matérn covariance is the limit as ν approaches infinity, giving the squared exponential covariance function

$$C(\boldsymbol{x}_1, \boldsymbol{x}_2; \sigma^2, r) = \sigma^2 \exp\left(-\frac{|\boldsymbol{x}_1 - \boldsymbol{x}_2|^2}{2r^2}\right).$$

The parameters r and σ^2 play the same role as in the exponential covariance model. Although superficially similar, these two covariance functions result in processes realizations with drastically different smoothness properties. The exponential covariance function results in extremely rough realizations, while the squared exponential covariance function results in extremely smooth realizations.

Figure 4 shows three different Matérn class covariance functions and a sample realization from each. The covariance functions all have $\sigma^2 = r = 1$ and differ only in ν. Shown are the exponential covariance, squared exponential covariance, and the $\nu = 1.5$ covariance. The dramatically different behavior of the exponential and squared exponential covariance is readily apparent. The covariance with $\nu = 1.5$ is something of a compromise: still quite smooth compared to the exponential model but permitting more local variation than the squared exponential.

Creating New Covariance Functions

Although the Matérn class of covariance functions are the ones most commonly seen in spatial statistics, there are many other useful covariance functions besides the ones discussed here. Moreover, covariance functions can be combined in interesting ways

to create new ones. Sums and products of covariance functions are also valid covariance functions, and this property makes it possible to design a Gaussian process model with significant structure. For example a time series could be modeled by combining a squared exponential covariance function representing a slow moving long-term trend with a periodic covariance function representing seasonal variation.

LEARNING COVARIANCE FUNCTION PARAMETERS

Up to now we have assumed that the covariance function parameters are known. This is unlikely to be the case in realistic modeling situations. In this section we will discuss and illustrate two approaches to learning covariance function parameters from data.

A fully Bayesian treatment of the hierarchical models we have examined involves assigning prior distributions to unknown parameters and updating these distributions with data to obtain posterior distributions. The prior distributions represent our uncertainty regarding the values of the parameters, while the posterior distributions are used for inference. This process is typically carried out by Markov chain Monte Carlo methods.

A different approach is to assign to unknown parameters values that maximize the marginal likelihood. A marginal likelihood is a likelihood function in which some parameter has been marginalized or "integrated out." This approach is known as empirical Bayes. In the context of Gaussian processes, this refers to marginalizing the random function from its joint distribution with the training data.

We will illustrate both techniques with a simple Gaussian process regression example on simulated data. The data is simulated from a mean zero Gaussian process

with a squared exponential covariance function with parameters $\sigma^2 = 1$ and $r = 1$. Additionally, a small amount of observation error generated from $N(0, \tau^2)$ is added to the training data, with $\tau^2 = 0.01$. The model can be written hierarchically as

$$y_i = f(x_i) + \epsilon_i$$

$$f \sim GP\big(0, C(x_1, x_2; \sigma^2, r)\big)$$

$$\epsilon \sim N(0, \tau^2),$$

where $C(x_1, x_2; \sigma^2, r)$ is the squared exponential covariance function.

Marginal Likelihood and Empirical Bayes

In this section we demonstrate the empirical Bayes approach to learning covariance function parameters. Recall that under noisy sampling the joint distribution of the n training outputs $\boldsymbol{y}$ and the random function $\boldsymbol{f}^*$ is

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{f}^* \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix} C_{nn} + \tau^2 I_n & C_{n*} \\ C_{*n} & C_{**} \end{pmatrix}\right).$$

Using the earlier result on multivariate normal marginal distributions we can obtain the marginal distribution $\boldsymbol{y} \sim N(0, C_{nn} + \tau^2 I_n)$. Therefore the log marginal likelihood is

$$\log \mathcal{L}(\sigma^2, r, \tau^2 | \boldsymbol{y}) = -\frac{1}{2}\boldsymbol{y}^T (C_{nn} + \tau^2 I_n)^{-1} \boldsymbol{y} - \frac{1}{2}\log|C_{nn} + \tau^2 I_n| - \frac{n}{2}\log 2\pi.$$

The parameters σ^2, r, and τ^2 can be found numerically by maximizing the log marginal likelihood. With the parameters known, we can proceed with noisy process prediction. This is illustrated in Figure 5.

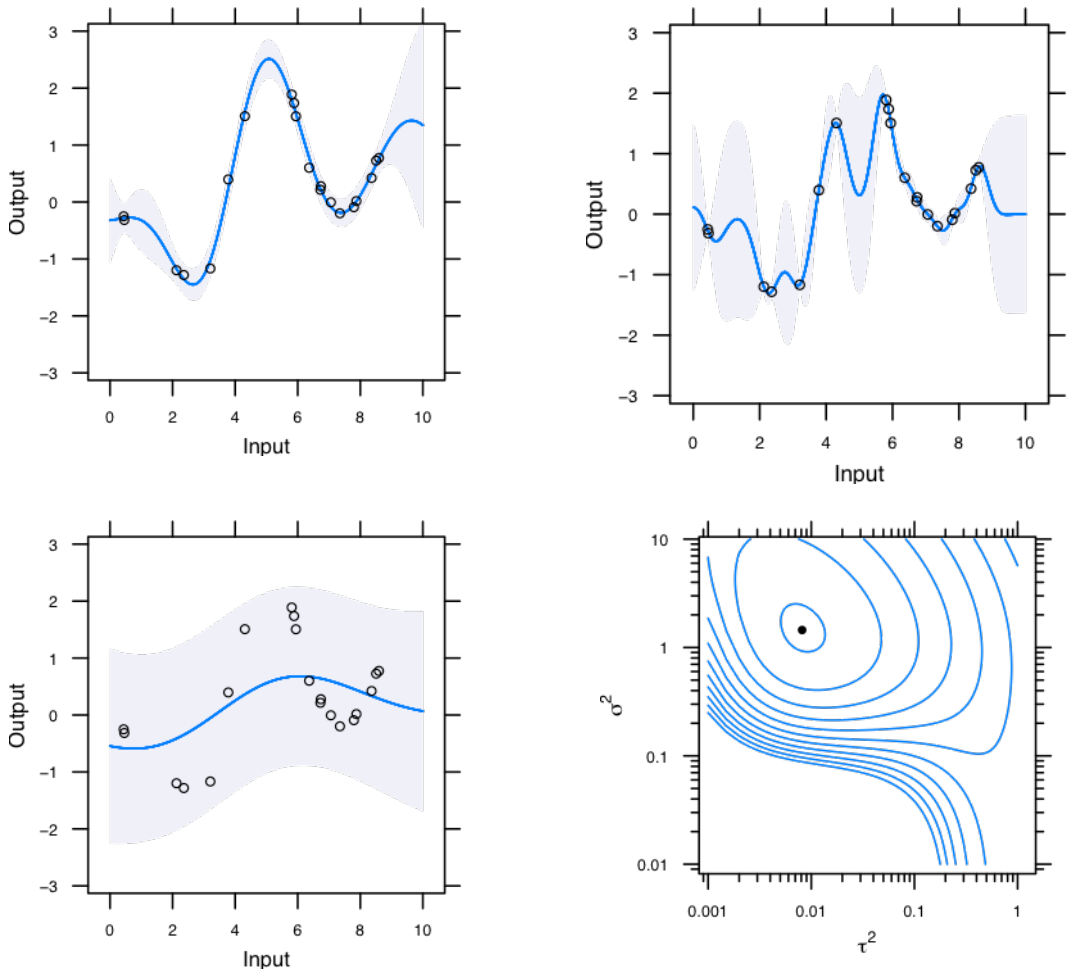

Figure 5: Learning covariance function parameters by empirical Bayes. The data was generated from a noisy Gaussian process with squared exponential covariance function. The upper left panel shows the process mean and prediction confidence band when the covariance function parameters are learned by maximizing the marginal likelihood. The upper right and lower left panels illustrate the effect of a too-short and too-long range parameter r, respectively. The lower right panels shows the log marginal likelihood contours when r is held to the true value.

The upper left panel shows the posterior process mean in blue and the predictive confidence interval calculated using the parameter values obtained by maximizing the marginal likelihood, $\sigma^2 = 2.035, r = 1.323$, and $\tau^2 = 0.011$. Notice that the confidence bands grow large at test locations far from the training points. In the upper right and lower left panels we show the result of process prediction when the covariance function parameters are learned by marginal likelihood maximization. For simplicity we fix r and obtain the marginal variance σ^2 and error variance τ^2 by optimization.

In the upper right panel the range parameter is fixed at $r = 1/3$, substantially smaller than the true value. The other parameters, as determined numerically, are $\sigma^2 = 0.7$ and $\tau^2 = 0.002$. The reduced range parameter has the effect of making the process wigglier, resulting in a mean that passes very nearly through the training data. The error variance has been reduced significantly, from $\tau^2 = 0.01$ to $\tau^2 = 0.002$. Notice however that the increased flexibility of the process results in much larger confidence bands away from the training data compared to the upper left panel. This is particularly evident near $x = 5$. In summary, the reduced range means that a quickly varying process with small observation error best explains the training data.

In the lower left panel we have the opposite situation with the range fixed at $r = 3$, much larger than the true value. The other parameters are $\sigma^2 = 0.6$ and $\tau^2 = 0.6$, so there is less overall variability but much greater observation error. The process is now fairly rigid and remains close to the overall mean of the observed data, and the

confidence bands have almost uniform width. In this case a slowly moving process with large error variance best explains the data. Effectively we are relying on observation error to explain the data.

The lower right panel shows log marginal likelihood contours when $r = 1$, the true value. These were obtained by evaluating the log marginal likelihood on a grid of σ^2 and τ^2 values. The covariance function parameters were obtained by seeking the maximum value on the grid.

Markov Chain Monte Carlo

The fully Bayesian approach requires prior distributions to be assigned to each unknown quantity. In this example we use:

$$\pi(\sigma^2) \sim IG(2,1)$$

$$\pi(r) \sim HalfCauchy(0,25)$$

$$\pi(\tau^2) \sim HalfCauchy(0,25)$$

The inverse gamma distribution $IG(\alpha, \beta)$ with shape parameter $\alpha = 2$ has infinite variance and is centered on the scale parameter β, in this case $\beta = 1$. The half-Cauchy distributions are uninformative but have mass near zero, as we do not wish to preclude small values for r and τ^2 in our posterior distributions.

Simulation was carried out using a simple random walk Metropolis scheme with each parameter σ^2, r, and τ^2 updated individually. These parameters all have positive support so sampling was done on a log scale with normal proposal distributions. Since the sampling was done on transformed variables the Jacobian determinant was included

in the joint posterior according to the law of transformation. The proposal distributions were tuned so that each parameter had an acceptance rate close to 0.45, in line with an approximately optimal rate for a normal target with a normal proposal. Two chains were run for 25000 iterations each with Initial values drawn from the prior distributions. The final 10000 samples from each chain were retained for inference.

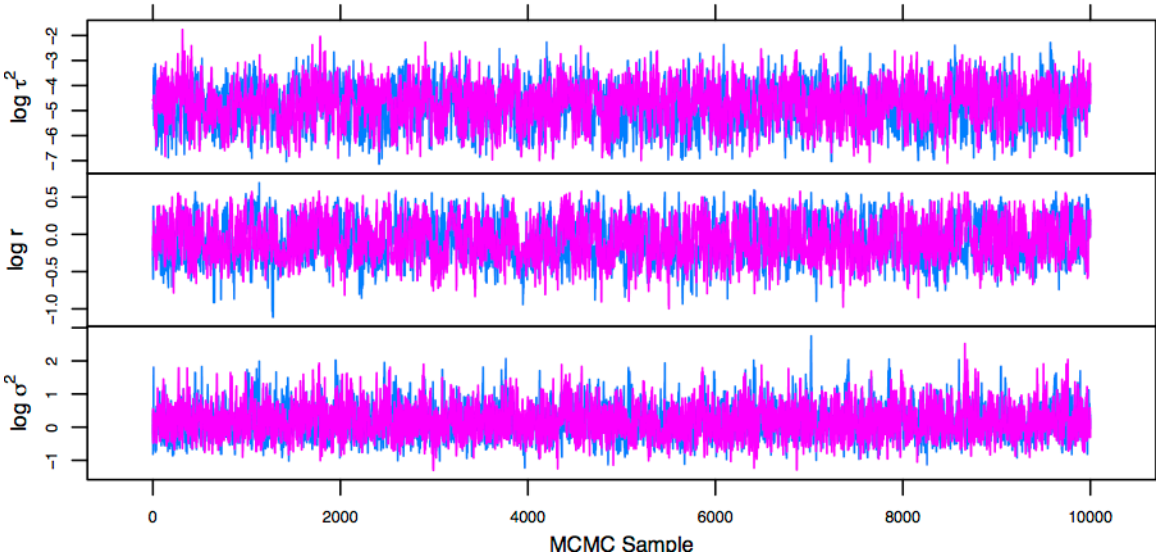

Figure 6: Learning covariance function parameters by MCMC. Two chains were run for 25000 iterations each and the final 10000 samples were retained for inference. These log scale trajectory plots show satisfactory mixing.

Figure 6 shows trajectory plots on the log scale. These plots show that each parameter mixes fairly well. The parameter densities (on the normal scale) are shown in Figure 7 and summarized in Table 1.

Parameter	*Mean*	*SD*	*2.5%*	*50%*	*97.5%*
σ^2 (marginal variance)	1.373	0.855	0.507	1.147	3.645
r (range)	0.952	0.269	0.563	0.880	1.529
τ^2 (error variance)	0.011	0.010	0.002	0.009	0.037

Table 1: Learning covariance function parameters by MCMC. Posterior inference summary.

For 1000 posterior samples $\{(\sigma^{2(t)}, r^{(t)}, \tau^{2(t)}) : t = 1, \ldots, 1000\}$ noisy prediction was carried out to obtain the latent function mean and pointwise prediction confidence intervals. The average mean function and average confidence band are displayed in Figure 8. The results are quite similar to those obtained by empirical Bayes, but these results account for the uncertainty in the values of the covariance function parameters and error variance.

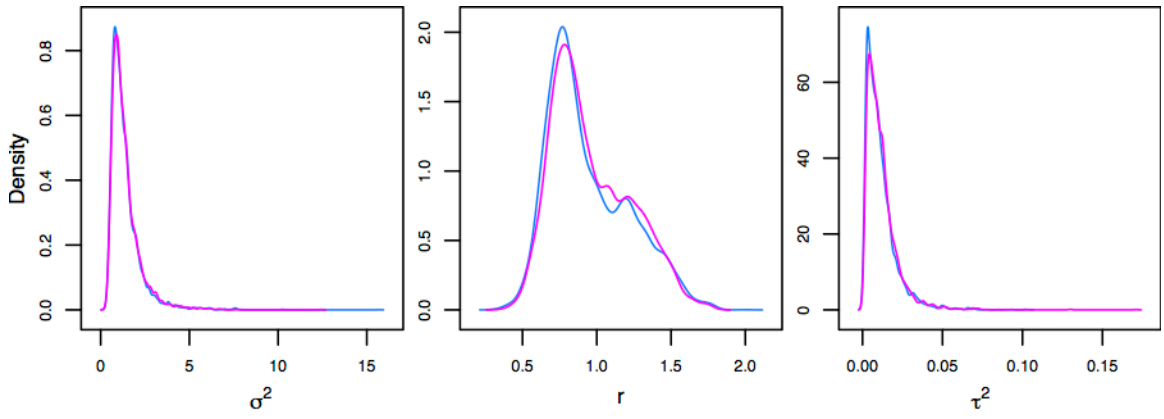

Figure 7: Learning covariance function parameters by MCMC. This figure shows posterior densities.

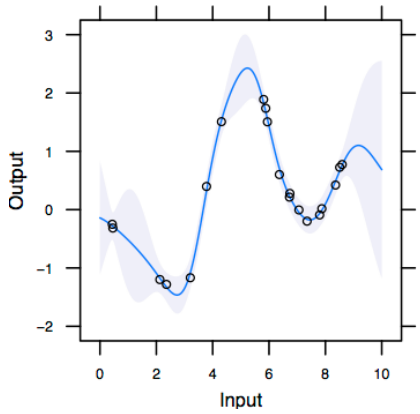

Figure 8: Learning covariance function parameters by MCMC. This figure shows the average process mean and average confidence band for 1000 parameter samples.

Spatially Varying Coefficients Model

The spatially varying coefficients model augments a linear regression with a spatial process to accommodate spatially varying regression coefficients. Allowing regression coefficients to vary spatially can account for spatial dependence in the outcome variable and in the relationship between the outcome variable and its covariates. In this section we will extend the ideas developed in the previous section to multivariate Gaussian processes and apply them to spatial modeling.

SPATIAL MODELING

There is now a rich literature on spatial statistics. Classical geostatistics focused on spatial interpolation by kriging, sometimes called optimal spatial prediction. This technique takes many forms but in essence it constructs an estimate of the covariance function from data and uses it to make spatial prediction as discussed above in Gaussian Process Predictive Distributions. (The primary object of interest in geostatistics is the variogram, which gives the variance of the difference between observations as a function of distance, or distance and direction in the case of anisotropic models. This is related to covariance in a straight-forward way.) The mean of the predictive distribution is taken to be the predicted value, and the variance is used to estimate the prediction uncertainty. However, in classical geostatistics these formulas are motivated by seeking minimum variance unbiased estimators rather than by analysis of Gaussian processes.

More recently, Monte Carlo methods have enabled the use of Bayesian hierarchical models in spatial prediction problems. These models seem to be natural in this context, but they do present computational challenges. Essentially this is because of

the requirement to invert the covariance matrix in each MCMC iteration. Matrix inversion is an $O(n^3)$ operation, meaning that the time required to invert an $n \times n$ matrix is proportional to n^3, so these models are limited to datasets with at most a few thousand observations. The problem is exacerbated in spatio-temporal models. Research on techniques for large datasets with tens of thousands or millions of observations is ongoing.

SPATIALLY VARYING COEFFICIENTS

When spatial locations are indexed by $\boldsymbol{s} \in D$, $D \subset \mathbb{R}^d$, for example by latitude and longitude, then the outcome variable $y(\boldsymbol{s})$ is modeled as

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta} + \boldsymbol{z}(\boldsymbol{s})^T\boldsymbol{w}(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),$$

where $\boldsymbol{x}(\boldsymbol{s})$ is a vector of p spatially indexed covariates, $\boldsymbol{\beta}$ is the associated vector of regression coefficients, $\boldsymbol{z}(\boldsymbol{s})$ is the $q \leq p$ dimensional subset of $\boldsymbol{x}(\boldsymbol{s})$ with spatially varying coefficients, and $\epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$. In the case that all of the covariates have spatially varying coefficients, that is, in the case that $\boldsymbol{z}(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})$, then the spatially varying coefficients are defined as $\widetilde{\boldsymbol{\beta}}(\boldsymbol{s}) = \boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s})$ and the model can be written as $y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\widetilde{\boldsymbol{\beta}}(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$. In other cases the spatially varying coefficients are defined in the same way but only for those covariates included in $\boldsymbol{z}(\boldsymbol{s})$. The residual variance τ^2 captures measurement error and micro-scale spatial variation.

The vector $\boldsymbol{w}(\boldsymbol{s})$ follows a mean zero multivariate Gaussian process. Thus we envision at each location $\boldsymbol{s}$ a multivariate normal random variable with mean zero and covariance matrix $C_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{s})$. Since the mean is known, the spatial process is completely

characterized by its matrix valued cross covariance function $C_{\boldsymbol{w}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$. Notationally, $\boldsymbol{w}(\boldsymbol{s}) \sim GP(\boldsymbol{0}, C_{\boldsymbol{w}}(\boldsymbol{s}_1, \boldsymbol{s}_2))$. Cross covariance functions are multivariate generalizations of the ordinary covariance functions associated with univariate Gaussian processes. The cross covariance function gives the covariance between components of the spatial process at locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, that is to say $\text{cov}\big(\boldsymbol{w}(\boldsymbol{s}_1), \boldsymbol{w}(\boldsymbol{s}_2)\big) = C_{\boldsymbol{w}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$. For n observations in D, this defines a covariance matrix $\Sigma_{\boldsymbol{w}}$ which is an $nq \times nq$ matrix partitioned into $q \times q$ blocks where block (i, j) is equal to $C_{\boldsymbol{w}}\big(\boldsymbol{s}_i, \boldsymbol{s}_j\big)$.

Cross covariance functions are rather more difficult to specify than the ordinary covariance functions associated with univariate Gaussian processes. In particular they must be specified with care because the resulting covariance matrix $\Sigma_{\boldsymbol{w}}$ is required to be symmetric positive definite. The symmetry requirement on $\Sigma_{\boldsymbol{w}}$ implies that at each pair of locations $\boldsymbol{s}_1 \in D$ and $\boldsymbol{s}_2 \in D$ the matrix $C_{\boldsymbol{w}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$ must satisfy $C_{\boldsymbol{w}}(\boldsymbol{s}_1, \boldsymbol{s}_2) = C_{\boldsymbol{w}}(\boldsymbol{s}_2, \boldsymbol{s}_1)^T$. Furthermore as $\boldsymbol{s}_2 \to \boldsymbol{s}_1$, $C_{\boldsymbol{\beta}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$ must become symmetric positive definite because it gives the covariance of components of the spatial process within site $\boldsymbol{s}_1$. Several different methods of constructing valid cross covariance functions are developed in the literature, including kernel convolution and multivariate correlation functions. Here we use the standard geostatistical technique known as the linear model of coregionalization. Under this model the multivariate spatial process $\boldsymbol{w}(\boldsymbol{s})$ arises from a linear transformation of independent unit variance spatial processes,$\boldsymbol{w}(\boldsymbol{s}) = L(\boldsymbol{s})\boldsymbol{e}(\boldsymbol{s})$, where the coregionalization matrix $T(\boldsymbol{s}) = L(\boldsymbol{s})L(\boldsymbol{s})^T$ is a $q \times q$ covariance matrix, $\boldsymbol{e}(\boldsymbol{s}) = \big(e_1(\boldsymbol{s}), \dots, e_q(\boldsymbol{s})\big)^T$, and each e_i, $i = 1, \dots, q$, is an independent unit variance

spatial process with covariance function $\rho_i(\boldsymbol{s}_1, \boldsymbol{s}_2)$. Notably, each $\rho_i(\boldsymbol{s}_1, \boldsymbol{s}_2)$ may have it's own parametric form, smoothness, and range properties. Although this offers considerable modeling flexibility, since each component of $\boldsymbol{w}(\boldsymbol{s})$ is a linear combination of the components of $\boldsymbol{e}(\boldsymbol{s})$, the smoothness of the components in the resulting multivariate process is dictated by the smoothness of the roughest component of $\boldsymbol{e}(\boldsymbol{s})$. This limitation can potentially be overcome by imposing structural zeros in the coregionalization matrix (that is, fixing certain entries to zero), but it has led some researchers to eschew the multivariate approach in favor of modeling each component separately. The special case of identical and independently distributed spatial processes, that is when every $\rho_i(\boldsymbol{s}_1, \boldsymbol{s}_2)$ is identical and $T(\boldsymbol{s}) = \sigma^2 I_q$, is called the separable case. Although less flexible, this form benefits from substantial computational simplifications.

Under the linear model of coregionalization, the resulting cross covariance for $\boldsymbol{w}(\boldsymbol{s})$ is

$$\begin{aligned} C_{\boldsymbol{w}}(\boldsymbol{s}_1, \boldsymbol{s}_2) &= \operatorname{cov}\big(L(\boldsymbol{s}_1)\boldsymbol{e}(\boldsymbol{s}_1), L(\boldsymbol{s}_2)\boldsymbol{e}(\boldsymbol{s}_2)\big) \\ &= L(\boldsymbol{s}_1)\boldsymbol{e}(\boldsymbol{s}_1)\big(L(\boldsymbol{s}_2)\boldsymbol{e}(\boldsymbol{s}_2)\big)^T \\ &= L(\boldsymbol{s}_1)C_{\boldsymbol{e}}(\boldsymbol{s}_1, \boldsymbol{s}_2)L(\boldsymbol{s}_2)^T, \end{aligned}$$

where $C_{\boldsymbol{e}}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \operatorname{diag}\Big(\rho_1(\boldsymbol{s}_1, \boldsymbol{s}_2), \ldots, \rho_q(\boldsymbol{s}_1, \boldsymbol{s}_2)\Big)$, the cross covariance for $\boldsymbol{e}(\boldsymbol{s})$. The matrix $C_{\boldsymbol{e}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$ is diagonal by the assumption of independent components for $\boldsymbol{e}(\boldsymbol{s})$. When $L(\boldsymbol{s}) = L$ is constant, this construction of $\boldsymbol{w}(\boldsymbol{s})$ is identical to the construction of a multivariate normal random variable with covariance LL^T from a linear transformation of independent standard normal random variables. In fact, $C_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{s}) = LL^T$ since $C_{\boldsymbol{e}}(\boldsymbol{s}, \boldsymbol{s}) =$

I_q by the assumption of independent unit variance processes $e_i(\boldsymbol{s})$. In this case, without loss of generality we can assume a lower triangular L by the uniqueness of the Cholesky decomposition.

For n observations under the linear model of coregionalization with a constant L, $\boldsymbol{w} = (\boldsymbol{w}(\boldsymbol{s}_1)^T, \dots, \boldsymbol{w}(\boldsymbol{s}_n)^T)^T$ follows a multivariate normal distribution $N(0, \Sigma_{\boldsymbol{w}})$. Given $\boldsymbol{\beta}$, τ^2, and $\Sigma_{\boldsymbol{w}}$ we can write the model in hierarchical form

$$\boldsymbol{y}|\boldsymbol{w} \sim N(X\boldsymbol{\beta} + Z\boldsymbol{w}, \tau^2 I_n)$$

$$\boldsymbol{w} \sim N(0, \Sigma_{\boldsymbol{w}})$$

where $X = (\boldsymbol{x}(\boldsymbol{s}_1)^T, \dots, \boldsymbol{x}(\boldsymbol{s}_n)^T)^T$ is the usual linear regression design matrix and $Z = \text{diag}(\boldsymbol{z}(\boldsymbol{s}_1)^T, \dots, \boldsymbol{z}(\boldsymbol{s}_n)^T)$ is a block diagonal matrix with $\boldsymbol{z}(\boldsymbol{s}_i)^T$ as the i-th block. Thus the joint distribution is

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{w} \end{pmatrix} \sim N\left(\begin{pmatrix} X\boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} Z\Sigma_{\boldsymbol{w}}Z^T + \tau^2 I_n & Z\Sigma_{\boldsymbol{w}} \\ \Sigma_{\boldsymbol{w}}^T Z^T & \Sigma_{\boldsymbol{w}} \end{pmatrix} \right).$$

The vector $\boldsymbol{w}$ can be thought of as a vector of random effects. Though unobserved, it is of principal scientific interest because it explains how the covariates locally affect the outcome variable. Having obtained the joint distribution of the observations $\boldsymbol{y}$ and the random effects $\boldsymbol{w}$ we are able to proceed with either an empirical Bayes analysis or MCMC. We will adopt an MCMC approach and complete the Bayesian model by assigning a prior distribution to each unknown quantity.

For $\boldsymbol{\beta}$ we use the customary flat prior $\pi(\boldsymbol{\beta}) = 1$. To the top-level error variance τ^2 we assign a half-Cauchy prior. For the coregionalization matrix T we model the

individual entries of L with log-normal priors on the diagonal entries and normal priors on the off-diagonal entries. Alternatively we could place an inverse-Wishart prior on T, but some authors report poor mixing with this approach. Finally, the underlying spatial processes $e_i(\boldsymbol{s})$ were assigned Gaussian process priors with exponential covariance functions. Thus $e_i \sim GP(0, \rho_i(\boldsymbol{s}_1, \boldsymbol{s}_2))$ with $\rho_i(\boldsymbol{s}_1, \boldsymbol{s}_2) = \exp(-\phi_i|\boldsymbol{s}_2 - \boldsymbol{s}_1|)$ for $i = 1, \ldots, q$. The spatial decay parameters ϕ_i can be assigned gamma priors. The Monte Carlo methods for fitted the spatially varying coefficients model are described in the next section. The sampler was written in C++ using the Armadillo C++ linear algebra library and interfaced with R via the RcppArmadillo package.

SAMPLER DETAILS

Letting $\boldsymbol{\theta} = \{\phi_1, \ldots, \phi_q, \tau^2, L\}$, the joint posterior distribution is

$$\pi_n(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{w}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) \times \pi(\boldsymbol{\beta}) \times N(\boldsymbol{w}|0, \Sigma_w) \times N(\boldsymbol{y}|X\boldsymbol{\beta} + Z\boldsymbol{w}, \tau^2 I_n).$$

In order to reduce model dimensions and speed up MCMC convergence it is preferable to marginalize the spatial process $\boldsymbol{w}$. This can easily be done using the earlier result on multivariate normal marginal distributions, obtaining (given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$)

$$\boldsymbol{y} \sim N(X\boldsymbol{\beta}, Z\Sigma_w Z^T + \tau^2 I_n).$$

With the spatial process marginalized the joint posterior becomes

$$\pi_n(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) \times \pi(\boldsymbol{\beta}) \times N(\boldsymbol{y}|X\boldsymbol{\beta}, Z\Sigma_w Z^T + \tau^2 I_n).$$

While it is possible to further reduce dimensionality by marginalizing $\boldsymbol{\beta}$, it is typically of much lower dimension than $\boldsymbol{w}$, which grows in dimension with the number of data

points$n$, and therefore of much less concern. Moreover, the posterior complete conditional distribution of $\boldsymbol{\beta}$ can be derived and updated with a simple Gibbs sampler.

With $\pi(\boldsymbol{\beta}) = 1$ the posterior complete conditional distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y} \sim N(Bb, B)$$

$$b = X^T(Z\Sigma_w Z^T + \tau^2 I_n)^{-1}\boldsymbol{y}$$

$$B^{-1} = X^T(Z\Sigma_w Z^T + \tau^2 I_n)^{-1}X$$

The remaining parameters $\boldsymbol{\theta}$ are updated with Metropolis steps. Those parameters with positive support are transformed and sampled on the log scale, with the Jacobian determinant included in the posterior so that the correct acceptance ratios are used.

Given posterior draws $\left\{\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}\right\}_{t=1}^{T}$ we can recover $\boldsymbol{w}^{(t)}$ via the conditional distribution

$$\boldsymbol{w}^{(t)}|\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{y} \sim N(Bb, B)$$

$$b = \frac{Z^T\left(\boldsymbol{y} - X\boldsymbol{\beta}^{(t)}\right)}{\tau^{2(t)}}$$

$$B^{-1} = \Sigma_w^{-1} + \frac{Z^T Z}{\tau^{2(t)}}$$

The draws $\left\{\boldsymbol{w}^{(t)}\right\}_{t=1}^{T}$ are used for inference about the local effect of covariates on the response variable at observed locations. With $\boldsymbol{w}$ in hand, we can sample the posterior predictive distribution

$$\pi_n(\hat{y}(\boldsymbol{s})|\boldsymbol{y}) = \int N(\hat{y}(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta} + \boldsymbol{z}(\boldsymbol{s})^T\boldsymbol{w}(\boldsymbol{s}), \tau^2) \times \pi(\boldsymbol{\beta}, \boldsymbol{w}, \tau^2|\boldsymbol{y}) d\boldsymbol{\beta} d\boldsymbol{w} d\tau^2$$

by drawing from

$$\hat{y}(\boldsymbol{s})^{(t)} \sim N(\boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}^{(t)} + \boldsymbol{z}(\boldsymbol{s})^T\boldsymbol{w}(\boldsymbol{s})^{(t)}, \tau^{2(t)}).$$

To make predictions at an unobserved locations $\boldsymbol{s}^*$ we must first obtain $\boldsymbol{w}(\boldsymbol{s}^*)$. The joint distribution of $\boldsymbol{w}(\boldsymbol{s}^*)$ and $\boldsymbol{w}$ is

$$\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w}(\boldsymbol{s}^*) \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix} \Sigma_w & C_w(\boldsymbol{s}, \boldsymbol{s}^*) \\ C_w(\boldsymbol{s}, \boldsymbol{s}^*)^T & C_w(\boldsymbol{s}^*, \boldsymbol{s}^*) \end{pmatrix}\right),$$

where $C_w(\boldsymbol{s}, \boldsymbol{s}^*)$ is an $nq \times q$ block matrix with $C_w(\boldsymbol{s}_i, \boldsymbol{s}^*)$ as the i-th $q \times q$ block, for $i = 1, \ldots, n$. Thus the conditional distribution of $\boldsymbol{w}(\boldsymbol{s}^*)$ given $\boldsymbol{w}$ is

$$\boldsymbol{w}(\boldsymbol{s}^*)|\boldsymbol{w} \sim N(C_w(\boldsymbol{s}, \boldsymbol{s}^*)^T\Sigma_w^{-1}\boldsymbol{w}, C_w(\boldsymbol{s}^*, \boldsymbol{s}^*) - C_w(\boldsymbol{s}, \boldsymbol{s}^*)^T\Sigma_w^{-1}C_w(\boldsymbol{s}, \boldsymbol{s}^*)).$$

Given the posterior draws we can simulate $\boldsymbol{w}(\boldsymbol{s}^*)^{(t)}$ using the above conditional distribution. Finally we can draw from the posterior predictive at $\boldsymbol{s}^*$

$$\hat{y}(\boldsymbol{s}^*)^{(t)} \sim N(\boldsymbol{x}(\boldsymbol{s}^*)^T\boldsymbol{\beta}^{(t)} + \boldsymbol{z}(\boldsymbol{s}^*)^T\boldsymbol{w}(\boldsymbol{s}^*)^{(t)}, \tau^{2(t)}).$$

These conditional and predictive distributions make it possible to construct maps of the local effect of covariates on the response variable.

APPLICATION TO SIMULATED DATA

In this section we apply the spatially varying coefficients model to a simulated dataset where the relationship between the response variable and its covariates is spatially dependent. This allows us to evaluate the behavior of the sampler and the predictive capability of the model.

We randomly select $n = 200$ points in the unit square, as shown in Figure 9, and generate an outcome variable $\boldsymbol{y}$ from the spatially varying coefficients model. The design matrix X is $n \times 2$, with the first column the intercept and the second column a covariate

drawn independently from a standard normal. The regression coefficients are $\boldsymbol{\beta} = (5,1)^T$ and the error variance is $\tau^2 = 1$. The spatial process $\boldsymbol{w}(\boldsymbol{s})$ is generated under the linear model of coregionalization with $T_{11} = 1.5$, $T_{22} = 1$, and $T_{12} = T_{21} = 0.3$. The underlying processes $e_1(\boldsymbol{s})$ and $e_2(\boldsymbol{s})$ are drawn from mean zero Gaussian processes having exponential covariance functions with range parameters $r_1 = 1$ and $r_2 = 0.7$, respectively. Of the 200 points, 50 were held back as test data.

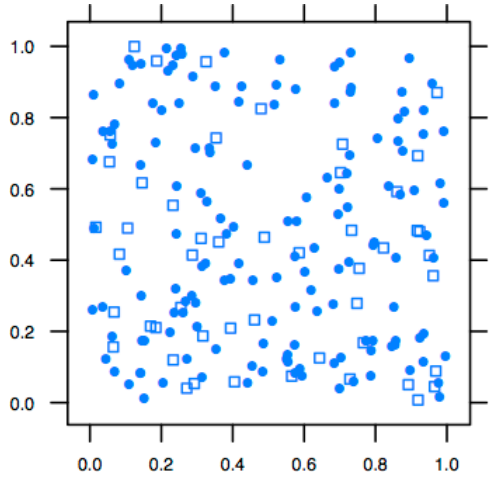

Figure 9: Observation locations for simulated data. Training points are indicated by filled circles, while test points are indicated by squares.

For the regression coefficients $\boldsymbol{\beta}$ we used the flat prior, $\pi(\boldsymbol{\beta}) = 1$. The error variance τ^2 was assigned a half-Cauchy prior with a large scale parameter, $\pi(\tau^2) = HC(0,25)$. The diagonal elements of L were assigned lognormal priors with location 0 and scale 1, $\pi(L_{11}) = \pi(L_{22}) = \text{lognormal}(0,1)$, while the off-diagonal element was assigned a standard normal prior, $\pi(L_{21}) = N(0,1)$. Finally, the spatial decay parameters

were assigned gamma priors, $\pi(\phi_1) = \pi(\phi_2) = \text{Gamma}(2,1)$. As before, the spatial process was marginalized to reduce model dimensions. The regression coefficients were sampled from their complete conditional distribution, while the remaining parameters were updated individually using Metropolis steps with normal proposals. In the case of parameters with constrained support (all but L_{21}), sampling was carried out on the log scale. The proposals were tuned to have a 0.45 acceptance rate.

Two chains with different initial values were run for 20000 iterations each. Convergence was assessed using Gelman-Rubin diagnostics, and the chains were found to have converged after approximately 5000 iterations. For inference, 500 samples were drawn randomly from the second half of each chain, yielding a total of 1000 samples. These are summarized in Table 2. Parameter trajectories and densities are shown in Figures 11 and 12. For each posterior sample $\left\{\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}\right) : t = 1, \ldots, 1000\right\}$ the posterior mean $\boldsymbol{w}$ was used to compute the posterior predictive mean for $\widehat{\boldsymbol{y}}$. Figure 10 shows a spline interpolation of the observed response $\boldsymbol{y}$ and the predictive mean $\widehat{\boldsymbol{y}}$, as well as spline interpolations of the spatially varying coefficients $\widetilde{\boldsymbol{\beta}}_0$ and $\widetilde{\boldsymbol{\beta}}_1$.

To evaluate the predictive capability of the model, we used the conditional distributions developed above to calculate the posterior mean response at the training and test locations. The mean absolute error (MAE) for m points, calculated as $m^{-1}\sum_{i=1}^{m}|y(\boldsymbol{s}_i) - \hat{y}(\boldsymbol{s}_i)|$, was 0.97 at training locations and 1.06 at test locations. The mean square error (MSE), calculated as $m^{-1}\sum_{i=1}^{m}\left(y(\boldsymbol{s}_i) - \hat{y}(\boldsymbol{s}_i)\right)^2$, was 0.94 at the

training locations and 1.68 at the test locations. These errors, as well as the smaller residual variance, improve significantly on a standard linear model as shown in Table 3.

Parameter	*True*	*Mean*	*SD*	*2.5%*	*25%*	*50%*	*75%*	*97.5%*
β_0 (intercept)	5.00	5.488	0.945	3.543	4.957	5.493	6.020	7.432
β_1 (covariate)	1.00	2.261	0.743	0.695	1.848	2.287	2.693	3.745
T_{11}	1.50	1.906	1.146	0.596	1.125	1.612	2.367	4.824
T_{22}	1.00	1.196	0.825	0.290	0.649	0.981	1.495	3.244
$T_{12}/\sqrt{T_{11}T_{22}}$	0.25	0.608	0.261	0.125	0.397	0.619	0.840	0.998
r_1 (intercept)	1.00	0.761	0.402	0.245	0.451	0.659	0.983	1.748
r_2 (covariate)	0.70	0.658	0.425	0.166	0.333	0.531	0.865	1.769
τ^2	1.00	1.340	0.235	0.910	1.181	1.325	1.485	1.831

Table 2: Posterior summaries of spatially varying coefficient model parameters for simulated data.

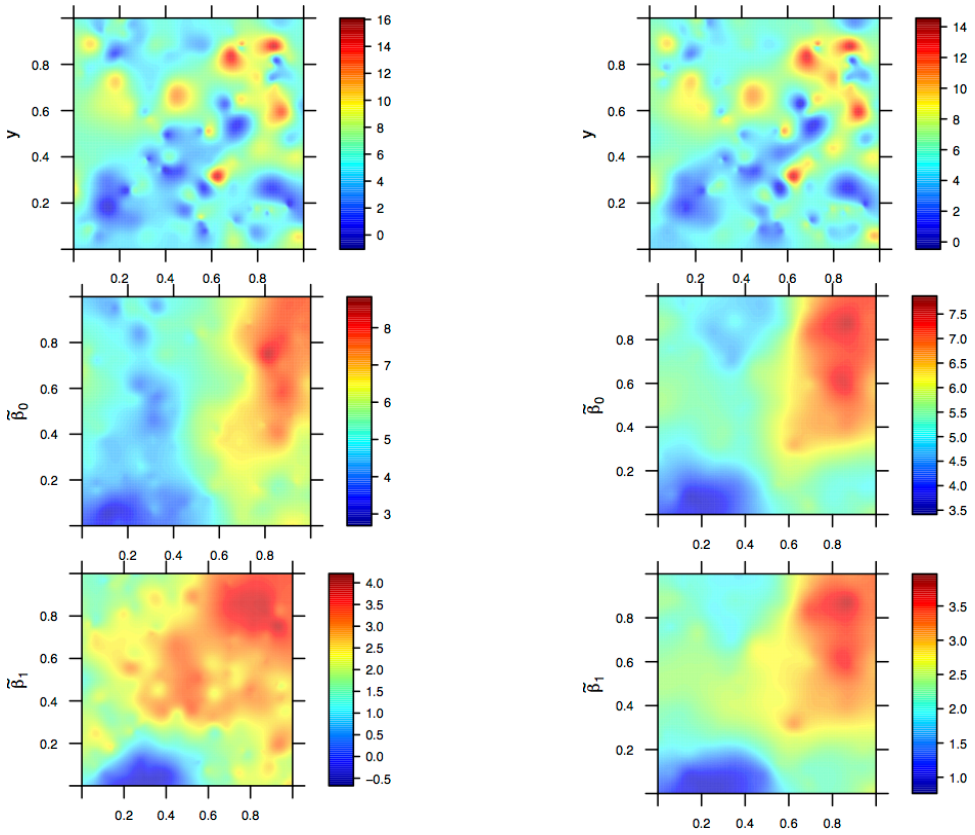

Figure 10: Spline interpolated surfaces calculated from the observed training data (left side) and posterior samples (right side). The top row shows the response variable. The middle and bottom rows show the spatially varying coefficients. The effect of the covariates on the response variable varies substantially across the domain.

Parameter	*True*	*Non-spatial*	*SVC*
β_0 (intercept)	5.00	5.539	5.493 (3.543, 7.432)
β_1 (covariate)	1.00	2.413	2.287 (0.695, 3.745)
T_{11}	1.50	-	1.612 (0.596, 4.824)
T_{22}	1.00	-	0.981 (0.290, 3.244)
$T_{12}/\sqrt{T_{11}T_{22}}$	0.25	-	0.619 (0.125, 0.998)
r_1 (intercept)	1.00	-	0.659 (0.245, 1.748)
r_2 (covariate)	0.70	-	0.531 (0.166, 1.769)
τ^2	1.00	3.034	1.325 (0.910, 1.831)
MAE		1.17	1.06
MSE		2.34	1.68

Table 3: Comparison of parameter values and prediction error for a linear model and the spatially varying coefficients model. The SVC parameters are reported as 50% (2.5%, 97.5%) percentiles.

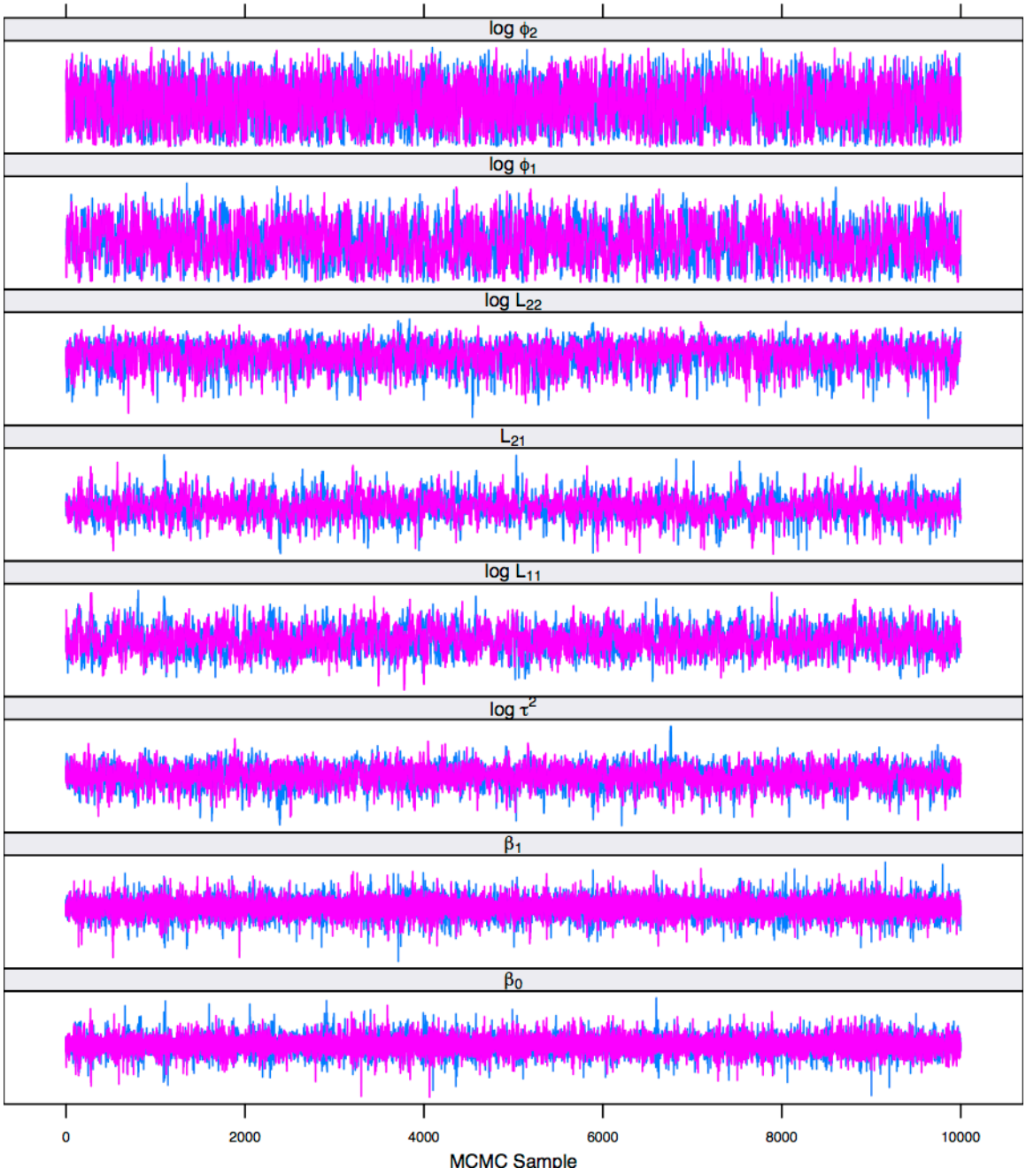

Figure 11: Two parallel chains were run for 20000 iterations and the final 10000 samples of each were retained for inference. These trajectory plots show satisfactory mixing.

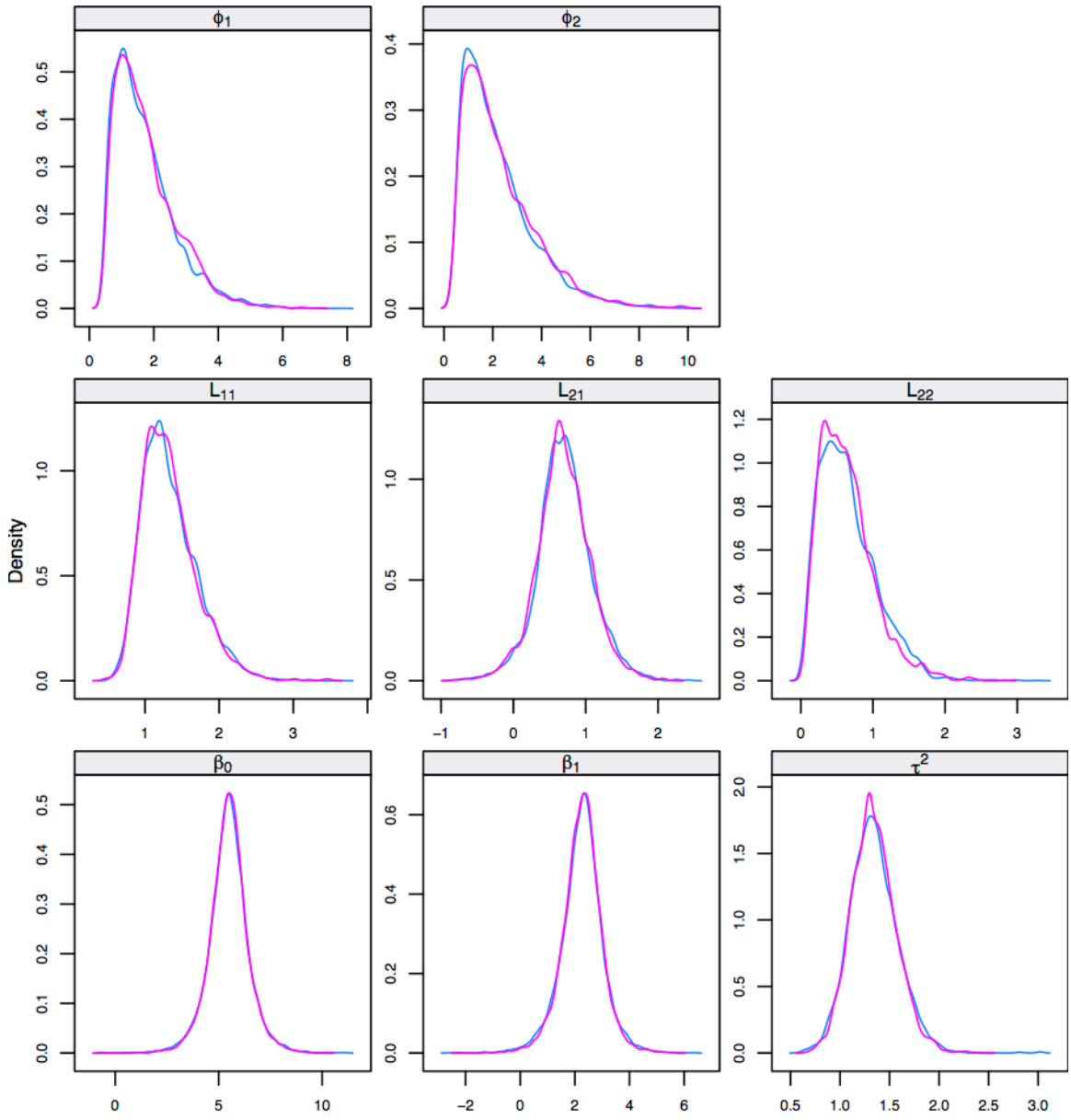

Figure 12: Posterior sample densities of model parameters for two parallel chains.

Discussion

In this report we developed the fundamentals of Gaussian process models beginning with the basic properties of the multivariate normal distribution. We demonstrated Gaussian process prediction when the process is observed exactly and when it is observed with error. We discussed the Matérn class of covariance functions and how covariance functions can be combined to introduce structure without resorting to parametric models. We used a simple Gaussian process regression problem to illustrate empirical Bayes and Markov chain Monte Carlo approaches to learning covariance functions, and pointed out connections to the spatial statistics literature.

We also extended Gaussian processes to a multivariate setting and introduced the geostatistical technique known as the linear method of coregionalization for constructing a matrix valued cross-covariance function. We applied these methods to a regression model in which regression coefficients are allowed to vary smoothly in space. This allows the covariates to affect the response variable differently in different regions, in contrast to an ordinary linear model in which the regression coefficients are fixed. We illustrated the technique with a case study on a simulated dataset, and found that the spatially varying coefficients model offered significant advantages over a linear model. The smaller prediction error combined with the smaller residual variance indicates a better fit without overfitting.

The primary advantage of the spatially varying coefficients model is that it is often quite natural to expect that the effect of a covariate on a response variable will differ from place to place or be correlated with the value of other covariates. For example

in a downscaling problem we might anticipate that a numerical weather simulation will be well calibrated in some geographic regions but not in others. If output from the simulation is used as a covariate with observed weather as a response, then allowing the regression coefficients to vary spatially can locally correct bias in the simulation. As another example we might expect the effect of elevation on temperature to be correlated with local humidity. Spatially varying regression coefficients allows complex relationships such as these to be modeled.

Another advantage of the spatially varying coefficients model is that it deals with spatial dependence in a systematic way through the covariance function. When applying a linear model to spatially referenced data, it is common to include some function of the coordinates as a covariate, for example latitude. However, this is quite arbitrary and raises many questions for the modeler. Should these coordinates enter in a linear fashion, or perhaps as a polynomial? If as a polynomial, then what degree? How should we interpret a regression coefficient associated with, say, latitude to the third power? The spatially varying coefficients model avoids these questions.

Several worthwhile extensions to the model are possible. It is possible to permit a non-Gaussian response by transforming a latent Gaussian variable, obtaining a spatially varying GLM. Another possibility is to allow a multivariate response. These would be useful for simultaneously modeling several related response variables. We could introduce a spatio-temporal version of the model by allowing the spatial processes to evolve in time. This could be done by treating model parameters as autoregressive processes as sampling them with the forward filtering backward sampling algorithm.

Finally, we could allow the coregionalization matrix to vary spatially so that the correlation between process and the spatial variance can adapt to local conditions.

References

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, *15*(2), 176–197. http://doi.org/10.1007/s13253-009-0004-z

Finley, A. O., Banerjee, S., Ek, A. R., & McRoberts, R. E. (2008). Bayesian multivariate process modeling for prediction of forest attributes. Journal of Agricultural, Biological, and Environmental Statistics, 13(1), 60–83. http://doi.org/10.1198/108571108X273160

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, *2*(2), 143–154. http://doi.org/10.1111/j.2041-210X.2010.00060.x

Gelfand, A., & Banerjee, S. (2010). Multivariate Spatial Process Models. *Handbook of Spatial Statistics*, 495–515. http://doi.org/doi:10.1201/9781420072884-c28

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–533. http://doi.org/10.1214/06-BA117A

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., & Banerjee, S. (2003). Spatial Modeling With Spatially Varying Coefficient Processes. *Journal of the American Statistical Association*, *98*(462), 387–396. http://doi.org/10.1198/016214503000170

Higdon, D. M. (2002). Space and space-time modeling using process convolutions. *Quantitative Methods for Current Environmental Issues*, 37–54. http://doi.org/10.1007/978-1-4471-0657-9_2

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. http://doi.org/10.1214/12-BA730

Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Sahu, S. K., Gelfand, a. E., & Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter, *11*(1), 61–86. http://doi.org/10.1198/108571106X95746

Schmidt, A. M. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research*, *108*(D24). http://doi.org/10.1029/2002JD002905